J. Clin. Chem. Clin. Biochem.
Vol. 18, 1980, pp. 433–437

Statistische Probleme beim Vergleich von klinisch-chemischen Analysen-Verfahren

Bericht über die Kleinkonferenz der Deutschen Gesellschaft für Klinische Chemie am 28. und 29. Juni 1979 in Hannover

Von *R. Haeckel*

(Eingegangen am 13. Mai 1980)

Teilnehmer/Participants:

Dr. *G. Bechtler*
c./o. Eppendorf Gerätebau GmbH
Barkhausenweg 1
2000 Hamburg 63

Prof. Dr. *K. Borner*
Freie Universität Berlin
Institut für Klinische Chemie
Hindenburgdamm 30
1000 Berlin 45

Prof. Dr. Dr. *J. Büttner*
Medizinische Hochschule Hannover
Institut für Klinische Chemie
Karl-Wiechert-Allee 9
3000 Hannover 61

Prof. Dr. *A. Delbrück*
Krankenhaus Oststadt
Institut für Klinische Chemie
Podbielskistraße 380
3000 Hannover 51

Prof. Dr. *U. Feldmann*
Abt. Medizinische Statistik, Biomathematik
und Informationsverarbeitung
Fakultät für Klinische Medizin Mannheim
der Universität Heidelberg
Theodor-Kutzer-Ufer
6800 Mannheim 1

Dr. *E. Hansert*
Max-Planck-Institut für Psachiatrie
Institut für Klinische Chemie
Kraepelinstraße 10
8000 München

Dr. *E. Henkel*
Krankenhaus Oststadt
Institut für Klinische Chemie
Podbielskistraße 380
3000 Hannover 51

Dr. *J. Kampmann*
Medizinische Hochschule Hannover
Med. Informatik
Karl-Wiechert-Allee 9
3000 Hannover 61

Prof. Dr. Dr. *H. Keller*
Kantonsspital
Klinisch-chemisches Zentrallabor
Frohbergstraße 3
9000 St. Gallen/Schweiz

PD Dr. *W. R. Külpmann*
Medizinische Hochschule Hannover
Institut für Klinische Chemie
Karl-Wiechert-Allee 9
3000 Hannover 61

Dr. *E. Markowetz*
c./o. Boehringer Mannheim
8132 Tutzing

Dipl.-Phys. *I. Mieth*
Medizinische Hochschule Hannover
Labordatenverarbeitung
Karl-Wiechert-Allee 9
3000 Hannover 61

Prof. Dr. *H.-J. Mitzkat*
Medizinische Hochschule Hannover
Innere Medizin
Karl-Wiechert-Allee 9
3000 Hannover 61

Dr. *E. Munz*
c./o. Boehringer Mannheim
8132 Tutzing

PD Dr. *M. Oellerich*
Medizinische Hochschule Hannover
Institut für Klinische Chemie
Karl-Wiechert-Allee 9
3000 Hannover 61

PD Dr. *A. J. Porth*
Medizinische Hochschule Hannover
Labordatenverarbeitung
Karl-Wiechert-Allee 9
3000 Hannover 61

Prof. Dr. *B. Schneider*
Medizinische Hochschule Hannover
Biometrie
Karl-Wiechert-Allee 9
3000 Hannover 61

Prof. Dr. Dr. *D. Stamm*
Max-Planck-Institut für Psychiatrie
Institut für Klinische Chemie
Kraepelinstraße 10
8000 München

PD Dr. *W. Vogt*
Universitätsklinik Großhadern
Institut für Klinische Chemie
8000 München

0340-076X/80/0018-0433$2.00

Organisation: R. Haeckel, Hannover

Zusammenfassung: Bisher publizierte Empfehlungen zur Evaluation von klinisch-chemischen Analysen-Verfahren beschrieben zwar das Versuchs-Protokoll detailliert, jedoch sind bei der statistischen Auswertung der Ergebnisse noch viele Fragen offen.

Die Kleinkonferenz hatte zum Ziel, Empfehlungen zu erarbeiten, welche statischen Verfahren insbesondere beim Methodenvergleich angewendet werden sollen. Diese Empfehlungen werden zur allgemeinen Diskussion gestellt.

Statistical problems in the comparison of methods of clinical chemical analysis

Report on the workshop conference of the German Society for Clinical Chemistry held on June 28 and 29, 1979 in Hannover

Summary: Despite published recommendations and experimental protocols for the evaluation of clinical chemical analytical methods, the statistical evaluation of results still leaves many questions unanswered.

The purpose of the workshop conference was to produce, and submit to discussion, recommendations for suitable statistical procedures, in particular for the comparison of different analytical methods.

Einführung

Bei der Evaluation von klinisch-chemischen Analysenverfahren (1–3) spielt der Methodenvergleich im allgemeinen eine zentrale Rolle. Während die bisher publizierten Empfehlungen das Versuchs-Protokoll detailliert beschreiben, sind bei der statistischen Auswertung der Ergebnisse noch viele Fragen offen. Daher wurde im Rahmen einer Kleinkonferenz versucht, Empfehlungen zu erarbeiten, welche statistische Verfahren bei der Evaluation von Analysenverfahren, insbesondere beim Methodenvergleich, angewendet werden sollten. Diese Empfehlungen sollen im Folgenden zur allgemeinen Diskussion gestellt werden.

Die Teilnehmer wollen diese Empfehlungen bei eigenen Evaluationsversuchen in Zukunft berücksichtigen, um praktische Erfahrungen zu sammeln, über die auf einer zweiten Kleinkonferenz zu einem späteren Zeitpunkt diskutiert werden soll. Gegebenenfalls werden dann die Empfehlungen nochmals überarbeitet und erneut publiziert.

Auf eine Zusammenfassung der einzelnen Referate wird verzichtet, da deren Darstellung wegen des vorwiegend mathematisch-statistischen Inhaltes nur in deren ganzer Länge allgemein verständlich wäre.

Im Folgenden wird die in den Richtlinien der Bundesärztekammer (4) angegebene Nomenklatur verwendet, entsprechende Empfehlungen der International Federation of Clinical Chemistry (3) in Klammern gesetzt. Die zugrunde gelegten Begriffsdefinitionen entsprechen den Vorschlägen der International Federation of Clinical Chemistry (3).

Untersuchung der Präzision

Grundsätzlich sollte beim Methoden-Vergleich zuerst die Präzision geprüft werden, da systematische Abweichungen nur erkannt werden können, wenn die zufälligen Fehler hinreichend klein sind.

Präzision in der Serie (within run imprecision)

Die Präzision in der Serie sollte sowohl mit nativen Human- als auch käuflichen Kontrollseren an drei verschiedenen Tagen bestimmt werden, und zwar mit mindestens drei verschiedenen Konzentrationen (unterer, mittlerer und oberer Meßbereich).

Die Berechnung der Standardabweichung, die zur Bestimmung der Präzision in der klinischen Chemie allgemein verwendet wird, erfolgt vorzugsweise mit 20 Werten, die hintereinander in einer Serie (d. h. in einem Segment ohne Zwischenkalibrierung) ermittelt werden. Zur Beurteilung der Präzision wird die Angabe von Mittelwert, Anzahl der Werte, der Standardabweichung und/oder des Variationskoeffizienten benötigt. Von den an den drei verschiedenen Tagen ermittelten Präzisionsdaten soll bei Publikationen der höchste und tiefste Wert jeweils angegeben werden.

Die Standardabweichung wird als Wurzel der Varianz nach der üblichen Formel (1)

$$s^2 = \frac{\sum_{i=1}^{i=n} (x_i - \bar{x})^2}{n - 1}$$

berechnet.

Diese Formel ergibt nur dann einen unverzerrten Schätzwert der Varianz, wenn die Meßwerte x_i der Serie alle denselben Erwartungswert μ und dieselbe Varianz σ^2 besitzen und statistisch unabhängig, also driftfrei und ohne Ausreißer sind. Solche systematischen Fehler können mittels einer Kontrollkarte der Residuen $(x_i - \bar{x})$ erkannt werden. Auf der Abszisse dieser Kontrollkarte wird die Positionsnummer i der Probe innerhalb der Serie und auf der Ordinate der zugehörige Wert des Residuums $(x_i - \bar{x})$ aufgetragen. Bei Konstanz des Erwartungswertes (und der Standardabweichung) schwanken die Residuen mit der vorgegebenen Wahrscheinlichkeit (z. B. 95% oder 99,75%) im Rahmen der To-

leranzgrenzen (z. B. ± 2σ oder ± 3σ-Grenzen). Die in den Richtlinien der Bundesärztekammer (4) angegebenen Beurteilungskriterien können hier ebenfalls angewendet werden.

Zur Identifizierung eines möglichen Ausreißers kann z. B. der r-Test eingesetzt werden (5, 6). Im übrigen gelten die späteren Ausführungen.

Präzision von Tag zu Tag (between-day imprecision)

Die Präzision von Tag zu Tag sollte, wie oben beschrieben, mit Kontrollseren an mindestens 10 Tagen in Doppelwerten zu Beginn der Serie bestimmt werden. Aus den zweiten Werten wird die Standardabweichung nach Formel 1 errechnet. Werden bei speziellen Analysenverfahren Mittelwerte aus Doppelbestimmungen verwendet, kann auch die Präzision von Tag zu Tag der Mittelwerte nach Formel 1 aus den Mittelwerten der Doppelbestimmungen berechnet werden. In diesem Fall sind täglich drei Werte hintereinander zu bestimmen und der jeweilige Mittelwert aus dem zweiten und dritten Wert zu ermitteln. Der erste Wert wird wegen eines eventuellen Verschleppungseffekts verworfen.

Verschleppungseffekte sollten unabhängig von der Präzision ermittelt werden.

Die Beurteilung der Präzision erfolgt im allgemeinen durch Vergleich mit der Referenzmethode (d. h. der zum Vergleich herangezogenen Methode). Andere Kriterien, die sich an der klinischen Relevanz orientieren, werden zur Zeit von verschiedenen Gruppen erarbeitet.

Vergleichende Untersuchungen mit Patientenproben

Wesentlicher Bestandteil der Richtigkeitsprüfung beim Methoden-Vergleich ist die möglichst gleichzeitige Untersuchung von mehreren Patientenproben (und eventuell Richtigkeitskontrollproben) mit der zu evaluierenden und einer Bezugsmethode, deren Auswahl und Deklaration nach den Empfehlungen der International Federation of Clinical Chemistry (3) erfolgen sollte. Die folgenden Versuchsanordnungen gelten für Serum- bzw. Plasmaproben und sind analog für andere Probenmaterialien zu modifizieren. Es sind mehrere Proben-Kollektive zu bilden, die unabhängig voneinander ausgewertet werden sollen. Bei den üblichen Methoden der Basis-Routine genügen Einfachbestimmungen pro Probe und Methode.

Das Haupt-Kollektiv besteht aus etwa 100 (10 Proben an je 10 Tagen) visuell unauffälligen Seren, deren Analyt-Konzentration möglichst gleichmäßig über einen großen Bereich (wenn möglich über den Meßbereich) verteilt ist. Bei speziellen Methoden kann die genannte Probenanzahl während der Versuchszeit unter Umständen nicht erreicht werden, so daß ein Kompromiß unvermeidlich ist. In manchen Fällen wird es sinnvoll sein, zwei oder drei Bereiche getrennt zu betrachten.

Weitere Kollektive sind bei Proben mit sichtbar enthaltenen Chromogenen (Hämoglobin, Bilirubin und Trübungen) sowie von bestimmten Patientengruppen zu bilden (z. B. Lebererkrankungen mit hohen Enzymaktivitäten, Niereninsuffizienz mit niedrigen Kreatinin-Clearance-Werten). Die Auswahl dieser Kollektive gilt vorwiegend für photometrische Verfahren. Bei anderen Meßprinzipien ist eventuell analog zu modifizieren.

Bei der Beurteilung der Ergebnisse lassen sich zwei verschiedene Fehlergruppen unterscheiden:

1. Proben-spezifische Fehler, die von Probe zu Probe erheblich schwanken können und die Streuung um die Ausgleichsgerade und deren Lage beeinflussen können (z. B. Interferenzen durch Medikamente).

2. Analysen-spezifische Fehler, die bei jeder Probenanalyse auftreten und entweder durch die Probenzusammensetzung oder Proben-unabhängig durch die Analysenverfahren bedingt sind. Bei diesen Fehlern handelt es sich entweder um konstante (Einfluß auf den Intercept der Ausgleichsgeraden), um proportionale Fehler (Einfluß auf die Steigung der Ausgleichsgeraden) oder um eine Kombination beider Fehler (z. B. Kalibrationsfehler).

Zur Differenzierung dieser Fehler wird folgendes Vorgehen empfohlen:

1. Der erste Schritt zur Beurteilung der Ergebnisse ist die graphische Darstellung im x/y-Koordinatensystem. Alle Werte sind einzuzeichnen, es sei denn, daß ein grober Bedienungsfehler identifiziert werden konnte. Als mögliche „Ausreißer" bezeichnete Werte, die nicht in die weiteren Berechnungen einbezogen wurden, sind durch ein entsprechendes Symbol auf der graphischen Darstellung zu kennzeichnen. Falls möglich, sollten solche „Ausreißer"-Analysen wiederholt und das Ergebnis der Wiederholungs-Analysen in einer Tabelle mitgeteilt werden.

Aus der graphischen Darstellung sollte optisch beurteilt werden, ob die Meßpunkte einer linearen Beziehung folgen. Ist keine Unlinearität zu erkennen, werden die nächsten Schritte vorgeschlagen.

2. Berechnung der Ausgleichsgeraden als standardisierte Hauptkomponente.

Die lineare Regression von y auf x oder x auf y sollte nur dann angewendet werden, wenn für eine der Variablen so präzise Werte ermittelt werden, daß sie praktisch wie Konstante behandelt werden können. Wenn dies beim Vergleich zweier Methoden nicht angenommen werden kann, wird ein multivariates Verfahren vorgeschlagen. Die standardisierte Hauptkomponenten-Analyse geht davon aus, daß beide Verfahren mit zufälligen Fehlern behaftet sind. Eine ausführliche Beschreibung findet sich bei l. c. (7).

Das von *Averdunk & Borner* (8) beschriebene Verfahren liefert dieselbe Ausgleichsgerade. Die standardisierte

Hauptkomponente entspricht dem geometrischen Mittel zwischen der Regression von y auf x und von x auf y. Die Steigung der standardisierten Hauptkomponente läßt sich aus der Streuung der x- und y-Werte einfach berechnen:

$$b = \frac{s_y}{s_x}$$

Ein weiterer Vorzug der Hauptkomponente ist, daß sie sich beim Vergleich mehrerer Methoden als structural relationship-Verfahren anwenden läßt (Tab. 1).

Tab. 1. Methoden zur Darstellung der linearen Beziehung zwischen Analysenverfahren.

1.	Vergleich zweier Verfahren
1.1	Regressionsbeziehung
1.2	Standardisierte Hauptkomponentenanalyse (orthogonale Regression)
2.	Vergleich mehrerer Verfahren
2.1	Linear structural relationship

3. Die berechnete Ausgleichsgerade wird nur innerhalb der Spannweite der graphisch dargestellten Wertepaare eingezeichnet und durch einen unterbrochenen Strich bis zur Ordinate verlängert. Ferner werden die Winkelhalbierende (y = x) und die Mittelwerte ($\bar{x}$ auf der Abzisse, $\bar{y}$ auf der Ordinate) eingezeichnet.

4. Vergleich der Mittelwerte mit Hilfe des t-Testes oder Vergleich der Verteilungsfunktion mit dem parameterfreien *Wilcoxon*-Test. Viele Statistiker halten den t-Test für ausreichend „robust" auch dann, wenn die Normalverteilung nicht gesichert ist.

5. Ebenso wird untersucht, ob sich die Steigung der standardisierten Hauptkomponente signifikant von 1,00 unterscheidet. Dies gilt für den Fall, daß bei beiden Methoden das gleiche numerische Resultat erwartet wird. Unterscheidet sich die Steigung signifikant von 1,00 liegt ein proportionaler Fehler vor. Als statistischer Test eignet sich der Likelihood Quotiententest (7) unter Annahme einer bivariaten Normalverteilung

$$LR = n \cdot \ln \frac{(s_1^2 - s_2^2)^2 + d^2}{d^2}$$

(wobei $d^2 = 4s_1^2 \cdot s_2^2(1 - r^2)$. Diese Testgröße ist χ^2-verteilt mit einem Freiheitsgrad.

6. Variable Fehler werden durch interferierende Komponenten mit von Probe zu Probe schwankenden Konzentrationen dann hervorgerufen, wenn beide zu vergleichende Methoden durch diese Interferenzen unterschiedlich beeinflußt werden. Solche variablen Fehler bewirken eine Streuung um die Ausgleichsgerade, die größer ist als die durch die zufälligen Fehler bei beiden Methoden bedingte Streuung.

Als Maß für die Streuung der Meßwerte um die Ausgleichsgerade haben *Westgard & Hunt* (9) den standard error of regression empfohlen, der nach *Cornbleet & Gochman* (10) wie folgt berechnet werden kann:

$$s_{y/x} = s_y \sqrt{\frac{n-1}{n-2}(1 - r^2)}$$

Der analoge Ausdruck für die standardisierte Hauptkomponente lautet (7):

$$s_{y/x} = s_y \sqrt{\frac{n}{n-1}(1 - |r|)}$$

7. Als weitere graphische Darstellungsmöglichkeiten werden über- bzw. nebeneinander projizierte Histogramme der Meßwerte oder der Residuen vorgeschlagen. Ferner können die Residuen auch in Abhängigkeit von der Probenpositionsfolge (Abzisse) dargestellt werden. Die letztere Darstellung eignet sich vor allem, um Abweichungen von der Linearität und Varianz-Inhomogenitäten zu erkennen.

Bei den Residuen kann als Bezugsgerade entweder die Winkelhalbierende y = x oder die Ausgleichsgerade (z. B. standardisierte Hauptkomponente) verwendet werden. Wenn von der Hypothese ausgegangen wird, daß beide Verfahren übereinstimmende Werte liefern, wird auf die Winkelhalbierende bezogen. Kann diese Hypothese aber nicht angenommen werden, bzw. wenn beide Methoden verschiedene Einheiten ergeben, sollte die Ausgleichsgerade als Bezugsmitte dienen.

Diese Verfahren dienen zur Beurteilung der Ergebnisse durch den Untersucher. Sie werden im allgemeinen nicht publiziert, es sei denn, daß ein besonderes Ergebnis demonstriert werden soll.

„Ausreißer"-Erkennung

Die Erkennung und Beurteilung von Ausreißern ist sehr komplex und konnte auf der Kleinkonferenz nur angeschnitten werden.

Um bereits während der Versuchs-Phase, also vor der endgültigen Auswertung eventuelle Ausreißer, die methodisch bedingt sind und nichts mit der Probe bzw. probeninhärenten systematischen Fehler zu tun haben, zu erkennen, sollen die täglich ermittelten Werte des Hauptkollektivs in einem x/y-Koordinatensystem dargestellt werden. Es wird die Winkelhalbierende y = x, sowie eine + 15%- und - 15%-Linie durch den Ursprung gezogen. Dieser Prozentsatz wurde als Beispiel angenommen und sollte individuell in Abhängigkeit von der analytischen Präzision festgelegt werden. Er dient als Warnbereich und sollte etwa dem Dreifachen des mittleren Variationskoeffizienten entsprechen (z. B. bei 3 VK = 5% eine obere Warngrenze von + 15%). Liegt ein Wert (erster Wert) außerhalb dieser Warngrenze, sollte die Analyse wiederholt werden (Ergebnis: 2. Wert). Es muß im Einzelfall entschieden werden, ob der erste Wert durch

den zweiten Wert ersetzt wird. Dies sollte im allgemeinen nicht erfolgen. Wird er aus der späteren Berechnung eliminiert, so muß er in der graphischen Darstellung gesondert als möglicher Ausreißer markiert werden. Er darf nur dann auch aus der graphischen Darstellung eliminiert werden, wenn ein grober bzw. systematischer Fehler identifiziert werden konnte.

Nach *Cornbleet & Gochman* sollten von der Berechnung der Hauptkomponente solche Werte ausgeschlossen werden, deren Residue (von der Ausgleichsgerade) das vierfache der Standardabweichung der Residuen (standard error of regression) überschreitet (10). Ein anderes Verfahren wurde kürzlich von *Healy* beschrieben (11).

Die auf der Kleinkonferenz erarbeiteten Vorschläge betreffen nur einige Aspekte eines umfassenden Methodenvergleichs. Ein umfassendes Protokoll zum Vergleich von Analysengeräten, das die hier publizierten Vorschläge berücksichtigen wird, wird zur Zeit von einer Arbeitsgruppe der Deutschen Gesellschaft für Klinische Chemie entwickelt.

Literatur

1. Barnett, R. N. & Youden, W. J. (1970), Amer. J. Clin. Pathol. *54*, 454–000.
2. Broughton, P. M. G., Gowenlock, A. H., McCormack, J. J. & Neill, D. W. (1974), Ann. Clin. Biochem. *11*, 207–218.
3. Büttner, J., Borth, R., Boutwell, J. H., Broughton, P. M. G. & Bowyer, R. C. (1976), this j. *14*, 265–275.
4. Richtlinien der Bundesärztekammer zur Durchführung von Maßnahmen der statistischen Qualitätskontrolle und von Ringversuchen im Bereich der Heilkunde (1970), Dt. Ärzteblatt *67*, 2228–2231.
5. Haeckel, R. (1975), Qualitätssicherung im medizinischen Labor, Deutscher Ärzteverlag GmbH, Köln.
6. Haeckel, R. & Schneider, B. (1980), GIT Labor-Medizin *3*, 81–84.
7. Feldmann, U., Schneider, B. & Haeckel, R. (1980), in Vorbereitung.
8. Averdunk, R. & Borner, K. (1970), this j. *8*, 263–268.
9. Westgard, J. O. & Hunt, M. R. (1973), Clin. Chem. *19*, 49–57.
10. Cornbleet, P. J. & Gochman, N. (1979), Clin. Chem. *25*, 432–438.
11. Healy, M. J. R. (1979), Clin. Chem. *25*, 675–677.

Prof. Dr. R. Haeckel
Karl-Wiechert-Allee 9
D-3000 Hannover 61

Willy Bürgi zum Honorar-Professor ernannt

Der Regierungsrat des Kantons Bern hat Herrn Dr. med. *Willy Bürgi*, Altpräsident der Schweizerischen Gesellschaft für Klinische Chemie, auf das Sommersemester 1980 zum Honorar-Professor für die Fächer Biochemie und Klinische Chemie an der Universität Bern ernannt.

Willy Bürgi ist neben seiner hauptamtlichen Tätigkeit als Chefarzt am Zentrallaboratorium des Kantonsspitals Aarau seit 1968 als externer wissenschaftlicher Mitarbeiter am Medizinisch-Chemischen Institut der Universität Bern tätig; dort ist er 1974 zum Privatdozenten ernannt worden. Sein vorbildlicher Einsatz hat wesentlich dazu beigetragen, daß trotz großer Jahresklassen von etwa 250 Studierenden im biochemischen Praktikum für die Vorkliniker ein Gruppenunterricht durchgeführt werden kann. Diese Tätigkeit, welche dank eines Entgegenkommens der Aargauischen Gesundheitsdirektion ermöglicht worden ist, sicherte eine enge Verbindung mit der Universität Bern, an der *Willy Bürgi* selbst sein Medizinstudium absolviert hat.

Mit der Ernennung zum Honorar-Professor hat nicht nur sein langjähriger Einsatz im Unterricht, sondern auch seine erfolgreiche wissenschaftliche Tätigkeit – vor allem zu Gunsten des Ausbaues der Qualitätskontrolle in der Klinischen Chemie in der Schweiz – die verdiente Anerkennung gefunden.